The relationship of Yellow card and Full-Time Away Team Goals*

Analysis of Primer league 2023-24 season

Che-Yu Wang

April 1, 2024

Football match outcomes, especially Full-Time Away Team Goals (FTAG), are difficult to predict in sports analytics. This study uses 2024 Premier League data to examine how Home Team Yellow Cards (HY) affect away team scoring. The study uses logistic, Poisson, and negative binomial regressions to find the best statistical model for football score count data. Poisson regression, ideal for count data, assumes equal mean and variance, which is often violated in real-world data. Logistic regression models binary outcomes and may overlook nuanced count data. Negative binomial regression better represents data variability by adding an overdispersion parameter. The study evaluates each model's ability to capture football scoring patterns' complexities using statistical analysis, model comparison, and diagnostic measures. The findings emphasize the delicate balance between model complexity and predictive accuracy, emphasizing the importance of choosing the right model to understand away team goal-scoring trends. This meticulous approach improves football analytics predictive modelling and sports statistics by refining count-based outcome analysis..

1 Introduction

You can and should cross-reference sections and sub-sections. We use R Core Team (2023) and Wickham et al. (2019).

The remainder of this paper is structured as follows. Section 2….

By 1900, a significant portion of the male population spent leisure time playing or watching sports, with football being a popular choice due to its spontaneity and order. The Victorian period marked the separation between work and leisure, with industrial work rhythms and

*Code and data are available at:https://github.com/Cheyuwang/Primer-league-analysis.

rapid urban growth causing a shift in behavior and promoting leisure (Mason 2023). The Premier League is the highest level of the English football league system. Contested by 20 clubs, it operates on a system of promotion and relegation with the English Football League.

The prediction of football match outcomes, particularly Full-Time Away Team Goals (FTAG), is a major challenge in the field of sports analytics. This challenge is driven not only by the inherent unpredictability of football matches, but also by a growing interest in understanding and forecasting sports events among academics and industry professionals. The ability to accurately predict these outcomes has important implications for strategic planning in sports management, the betting industry, and fan engagement. Despite the abundance of research in this area, football's dynamic and complex nature, characterized by unpredictable gameplay and a plethora of influencing factors, presents a persistent challenge in achieving high prediction accuracy. This gap emphasizes the importance of continuously refining predictive models capable of handling the complexities of football data.

This research uses a large-scale dataset from the football-data.co.uk website and focuses on the 2024 Premier League season. Several metrics are included in the dataset, such as Home Team Yellow Cards (HY), which are thought to affect the away team's capacity for scoring. In the context of football match outcomes, the study attempts to analyze and comprehend the predictive power of logistic, Poisson, and negative binomial regression models. Every model offers a unique method for managing count data: logistic regression models binary outcomes, Poisson regression handles count data with equal variance and mean, and negative binomial regression adds a parameter to handle overdispersion.

The comparative analysis of the study shows that, although logistic regression is a straightforward technique for forecasting discrete events, it might not make full use of the count data that is available, which could lead to information underutilization. Although poisson regression works well with count data, it often breaks down because it assumes equal mean and variance, which is rarely the case with football match results. Among the models that can better explain the variability and over dispersion found in football scoring data is negative binomial regression. The results indicate that negative binomial regression has a great deal of potential for increasing the accuracy of football match outcome predictions because they highlight the delicate balance between model complexity and predictive performance.

This study's contribution to the developing field of sports analytics—which offers knowledge that can greatly enhance football match prediction modeling—makes it noteworthy. The study is a useful tool for analysts, researchers, and practitioners of sports analytics because it identifies the best statistical model for FTAG prediction. The study also lays the foundation for future research by highlighting the significance of model selection in comprehending and predicting athletic events. The structure of the paper is as follows: After the introduction, Section 2 reviews the literature and provides an overview of the field's prior research as well as its theoretical foundations. The methodology—which covers the dataset, variables taken into account, and statistical models employed—is covered in Section 3. The results and analysis, which contrast the predictive powers of logistic, Poisson, and negative binomial regressions, are presented in Section 4. The implications of the results, the limitations of the study, and potential future research directions are covered in Section 5's conclusion.

2 Data

Our data is of penguins (Figure 1), from Horst, Hill, and Gorman (2020).

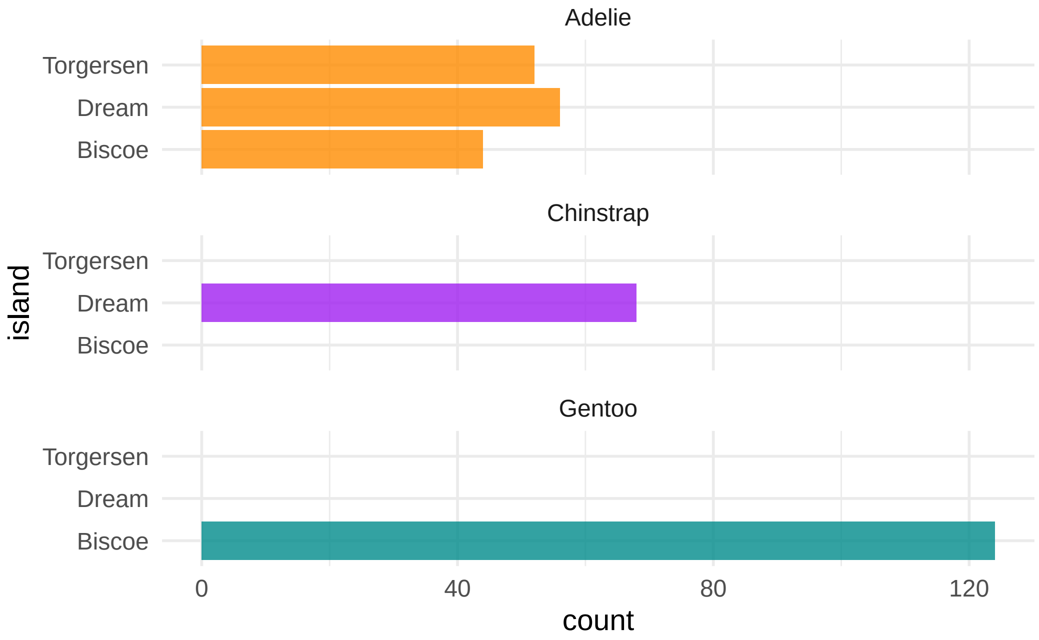

Figure 1: Bills of penguins

Talk more about it.

And also planes (**?@fig-planes**). (You can change the height and width, but don't worry about doing that until you have finished every other aspect of the paper - Quarto will try to make it look nice and the defaults usually work well once you have enough text.)

Talk way more about it.

3 Model

The goal of our modelling strategy is twofold. Firstly,…

Here we briefly describe the Bayesian analysis model used to investigate… Background details and diagnostics are included in Appendix B.

3.1 Model set-up

Define y_i as the number of seconds that the plane remained aloft. Then β_i is the wing width and γ_i is the wing length, both measured in millimeters.

$$\begin{aligned} y_i|\mu_i, \sigma &\sim \text{Normal}(\mu_i, \sigma) && (1)\\ \mu_i &= \alpha + \beta_i + \gamma_i && (2)\\ \alpha &\sim \text{Normal}(0, 2.5) && (3)\\ \beta &\sim \text{Normal}(0, 2.5) && (4)\\ \gamma &\sim \text{Normal}(0, 2.5) && (5)\\ \sigma &\sim \text{Exponential}(1) && (6) \end{aligned}$$

We run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`.

3.1.1 Model justification

We expect a positive relationship between the size of the wings and time spent aloft. In particular…

We can use maths by including latex between dollar signs, for instance θ.

4 Results

Our results are summarized in Table 1.

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

Table 1: Explanatory models of flight time based on wing width and wing length

	First model
(Intercept)	1.12
	(1.70)
length	0.01
	(0.01)
width	−0.01
	(0.02)
Num.Obs.	19
R2	0.320
R2 Adj.	0.019
Log.Lik.	−18.128
ELPD	−21.6
ELPD s.e.	2.1
LOOIC	43.2
LOOIC s.e.	4.3
WAIC	42.7
RMSE	0.60

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows...

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows...

References

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

Horst, Allison Marie, Alison Presmanes Hill, and Kristen B Gorman. 2020. *Palmerpenguins: Palmer Archipelago (Antarctica) Penguin Data*. https://doi.org/10.5281/zenodo.3960218.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.